Systems & Control Transactions

Research Article
Foundations of Computer Aided Process Design (FOCAPD 2024)
Breckenridge, Colorado, USA. July 14-18, 2024
Peer Reviewed Conference Proceeding

PSE PRESS

Development of Mass/Energy Constrained Sparse Bayesian Surrogate Models from Noisy Data

Samuel Adeyemo and Debangsu Bhattacharyya*

West Virginia University, Department of Chemical and Biomedical Engineering, Morgantown, West Virginia, USA
* Corresponding Author: Debangsu.Bhattacharyya@mail.wvu.edu.

ABSTRACT

This paper presents an algorithm for developing sparse surrogate models that satisfy mass/energy conservation even when the training data are noisy and violate the conservation laws. In the first step, we employ the Bayesian Identification of Dynamic Sparse Algebraic Model (BIDSAM) algorithm proposed in our previous work to obtain a set of hierarchically ranked sparse models which approximate system behaviors with linear combinations of a set of well-defined basis functions. Although the model building algorithm was shown to be robust to noisy data, conservation laws may not be satisfied by the surrogate models. In this work we propose an algorithm that augments a data reconciliation step with the BIDSAM model for satisfaction of conservation laws. This method relies only on known boundary conditions and hence is generic for any chemical system. Two case studies are considered-one focused on mass conservation and another on energy conservation. Results show that models with minimum bias are built by using the developed algorithm while exactly satisfying the conservation laws for all data.

Keywords: System Identification, Machine Learning, Algorithms, Design Under Uncertainty, Optimization

1.0 INTRODUCTION

Surrogate models are of great use for many process systems when developing the first-principles model is complex and/or time-consuming or the repeated simulation of the first-principles model for optimization/control is computationally expensive and/or difficult to converge reliably [1-3]. In recent years, greater access to numerous sensors, ease of collection and storage of large amount of data for many process systems and increased computational power with the emergence of supercomputers incentivized the development of data-driven models.

In the area of data-driven models, there are significant works on artificial neural network (ANN) models, with many powerful tools being readily available in the public domain. However, due to the black box nature, ANN models suffer from the lack of model interpretability which is a desired property of surrogate modeling. In addition, ANN models often require large amount of data for training while the developed models are often characterised with limited extrapolation capabilities. One approach to address these limitations is to develop data-driven interpretable models by using well-defined basis functions such as those employed in Sparse Identification of Nonlinear Dynamics (SINDy) [4], Automatic Learning of Algebraic Models (ALAMO) [5] and Algebraic learning via elastic net (ALVEN)[6]. However, these existing approaches can perform poorly when trained with the noisy data especially for the case of correlated noise which is common for the industrial data. For the SINDy algorithm in particular, several methods including ensemble modeling [7] and implicit SINDy [8] have been proposed to enhance the robustness of the algorithm to the noisy data. Still, the level of noise that is acceptable for satisfactory model building can be highly limited. In addition, these approaches often result in more complicated models leading to some loss of desired model interpretability and sparsity. In our previous work [9], we proposed the Bayesian identification of Dynamic Sparse Algebraic Models (BIDSAM) algorithm that addresses the challenges of sparsity, model interpretability and robustness to noisy data. The developed algorithm employs Bayesian inferencing implemented in the expectation maximization (EM) framework for simultaneous model parameter estimation and uncertainty quantification that

explicitly accounts for correlation in the measurement noise, if exists. In that algorithm, system nonlinearities are approximated by linear combinations of linear and nonlinear basis functions that are transformations of input variables and their interactive effects among themselves and the output variables. The optimal sparsest subset of the resulting large family of basis functions is selected using the branch and bound algorithm which returns a set of hierarchically ranked sparse models by minimizing an information criterion that not only incentivizes model fitness but also accounts for parameter estimability.

It is desired that models of chemical systems satisfy mass/energy constraints. Else the predicted results by a surrogate model can be meaningless. To this end, there exists several works in which physics-informed machine learning algorithms are developed for satisfying some physics constraints [10-12]. Most of these are not for satisfying mass or energy constraints, but for satisfying other physical constraints such as thermodynamic constraints. In addition, these approaches are highly specific requiring detailed knowledge of the system, which may not be available for a system where data-driven models are being developed.

In this work, we propose an approach that extends the capabilities of the BIDSAM algorithm for identification of robust sparse models that satisfy mass/energy conservation laws exactly both for the forward and inverse problems without loss of sparsity and model interpretability. The proposed approach guarantees the satisfaction of the conservation laws for all predictions while relying only on the knowledge of boundary conditions.

2.0 THEORY

2.1 Sparse Model Selection and Parameter Estimation

Consider a general nonlinear system with states x, inputs u, measured outputs y, and model parameters θ represented by the following set of differential and algebraic equations:

$$\dot{x} = f(x, u, \theta) \tag{1}$$

$$y = g(x) \tag{2}$$

The BIDSAM algorithm proposed in our previous work [9] approximates the system nonlinearities with a linear combination of linear and nonlinear transformations of input variables and their interactive effects among themselves and the measurements. The resulting model takes the following form:

$$y_k = \boldsymbol{A} y_{k-1} + \boldsymbol{C} U^*_{k-1} + \omega_k \tag{3}$$

where U^*_{k-1} represents the set of basis functions, ω_k stands for additive noise and $\boldsymbol{A}, \boldsymbol{C}$ are the model parameters. Using Bayesian inferencing implemented in the Expectation-Maximization (EM) framework, uncertainty quantification is undertaken while estimating model parameters. A branch and bound algorithm is developed to search for the sparsest subset of basis functions that best approximates the system behavior given the available data. The BIDSAM algorithm returns desired numbers of hierarchically ranked models based on a modified Akaike information criteria that is used as the model fitness criterion.

2.1 Model Update for Satisfaction of Conservation Laws

The models resulting from the BIDSAM algorithm show superior performance compared to existing algorithms especially when training data is corrupted with correlated noise. But the BIDSAM models are not guaranteed to satisfy conservation laws. This becomes very important especially when the training data are significantly noisy and violate the conservation laws. In this work, the BIDSAM model is augmented with a dynamic data reconciliation (DDR) step and an algorithm for updating model parameters so that final model results satisfy the mass/energy conservation laws. The overall model structure is given in Figure 1. The approach includes a linear transformation of the output from the BIDSAM model, x, by introducing a transformation that brings in more degrees of freedom for the optimization.

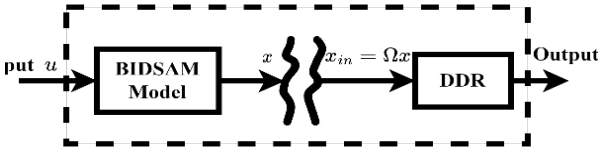

Figure 1: Mass/energy constrained model

2.2.1 Inverse problem

The inverse problem involves updating the BIDSAM model parameters in an iterative manner while solving the following optimization problem.

$$\min_{\Omega, y_{ddr}} \sum_i (\Omega x^k - y_{ddr})^T W_1^{-1} (\Omega x^k - y_{ddr}) + \sum_i (y_{data} - y_{ddr})^T W_2^{-1} (y_{data} - y_{ddr}) \tag{4}$$

$$s.t \text{ Mass/Energy balances}$$

In Equation 4, W_i's represent the weighting matrices, y_{data} denote the measured data and y_{ddr} denote the final model output that are desired to satisfy mass/energy balances. For the case of mass conservation, atom balance can take the following general form:

$$\sum n_{j,p_i}(u_p(k)) = \sum n_{j,q_o}(y_{q,ddr}(k)) \quad (5)$$

where n_{j,p_i} is the number of atoms of element j with respect to input variable p, and n_{j,q_o} are the number of atoms of element j with respect to output variable q. For energy conservation, a similar approach is followed by considering enthalpy in and out for the system as discussed later in section 3.2 for the second case study. Although equation 5 has been written for a general case of multiple component situation, the same conditions apply to scenarios in which only total mass flow is known at the inlet or outlet of the system. The situation in which only total flows in and out of the system are known is considered trivial and mass is assumed to be generally conserved conditioned on the known mass flow rate into the system as shown in the second case study.

The overall algorithm for solving the inverse problem is shown in Figure 2.

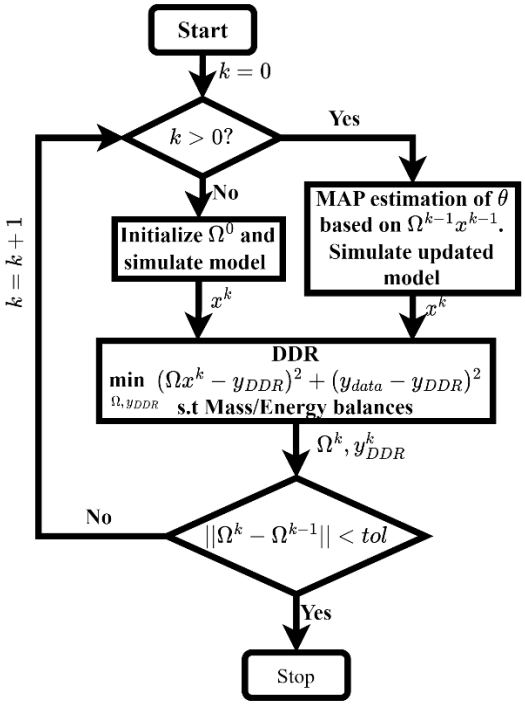

Figure 2. Sequential algorithm for inverse problem.

In the first iteration, initial values are assumed for the hyperparameters Ω, then we solve the optimization problem in Equation 4 using the Interior point (IPOPT) algorithm [13]. In subsequent iterations, model parameters are updated based on the optimal values obtained for Ω in the previous iteration. This is done by undertaking a maximum a posteriori (MAP) estimates of these parameters using Bayesian inferencing and the EM algorithm as detailed in our previous work [9]. The iteration continues till convergence.

2.2.2 Forward problem

The overall structure in Figure 1 remains the same for the forward problem. However, in this case, the parameters Ω and θ assume the optimal values obtained from the inverse problem and remain unchanged while DDR is undertaken just one time for each set of desired prediction. The following objective function replaces that in Equation 4 for the forward problem:

$$\min_{\Omega, y_{ddr}} \sum_i (\Omega x^k - y_{ddr})^T W_1^{-1} (\Omega x^k - y_{ddr}) \quad (6)$$

The constraints for this optimization remain the same as for the inverse problem.

3.0 CASE STUDIES

The efficiency of the developed algorithm is tested on two case studies: an isothermal CSTR and a lumped parameter superheater system.

3.1 Isothermal CSTR

For this case study, the performance of the developed algorithm is examined for the satisfaction of mass conservation. A schematic of the CSTR is shown in Figure 3. Chemical reactions for this system are as follows:

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C$$
$$2A \xrightarrow{k_3} D \quad (7)$$

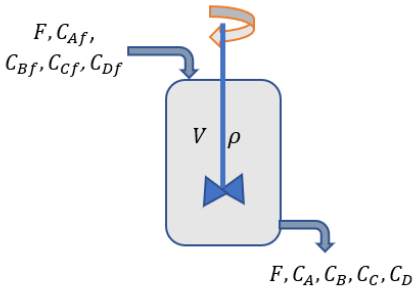

Figure 3. Schematic diagram of isothermal CSTR.

For this modeling purpose, the input variables considered are the reactor feed rate (F) and species concentrations at the inlet (C_{Af}, C_{Bf}, C_{Cf} and C_{Df}). We seek to predict the concentration of each species (C_A, C_B, C_C and C_D) at the reactor outlet based on changes in the input variables. Training data are simulated by solving the differential equations used to model the system from first-principles. Simulated correlated noise and constant bias

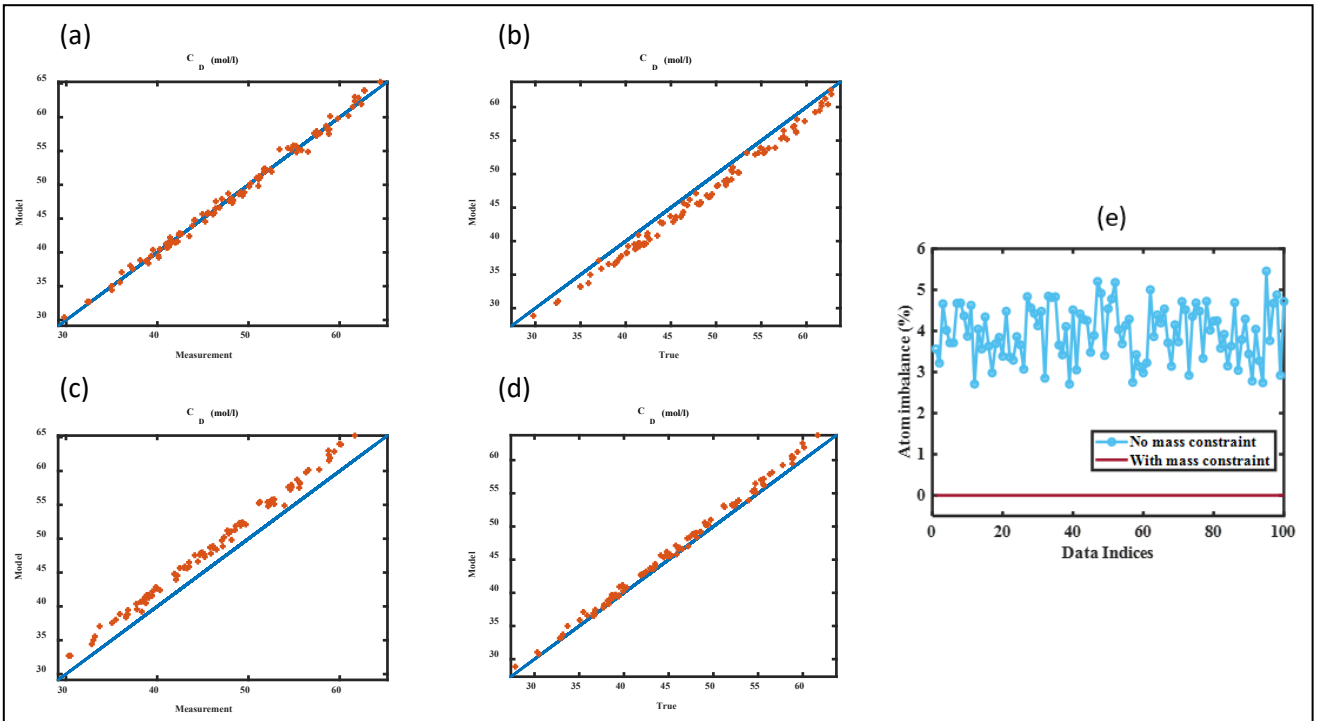

Figure 4. Validation plots for isothermal CSTR model: (a)-(b) BIDSAM, (c)-(d) BIDSAM with mass constraint.

Table 1. Performance measures of the models of the isothermal CSTR.

Mass constrained?	No. of parameters	% RMSE			
		C_A	C_B	C_C	C_D
Trained using True Data					
No	24	1.32	0.64	1.47	1.59
Yes		1.44	0.75	1.41	1.02
Trained using Noisy Data with Bias					
No	24	4.36	4.04	4.45	4.25
Yes		2.05	1.81	2.00	2.12

are added to this data which is then used for model training. Validation plots for a model obtained when training is done by using noisy data with bias are shown in Figure 4.

Figures 4(a-b) show that while the results from the BIDSAM model without constraints match the training data quite well but there is bias with respect to the truth. Figures 4(c-d) show that the results from the model built by using the algorithm with mass constraint exhibit bias with respect to the biased noisy data, but are closer to the truth. Figure 4(e) shows that while the model built without considering mass conservation constraints violates the mass balance, the algorithm with mass constraint included results in exact satisfaction of the conservation law. For training with noisy data with bias, the maximum %RMSE for the unconstrained algorithm is 4.45% while for the constrained algorithm we have 2.05%.

3.2 Adiabatic superheater system

This case study evaluates the performance of the BIDSAM model with energy constraints included. A superheater from a power plant, shown in Figure 5, where steam is superheated by using the hot flue gas is considered. The system has a cross-flow configuration. The input variables considered are the steam and flue gas mass flowrates $(\dot{m}_{st}, \dot{m}_{fg})$, temperatures $(T_{st,in}, T_{fg,in})$ and pressures $(P_{st,in}, P_{fg,in})$. The output variables are the steam and flue gas outlet temperatures $(T_{st,out}, T_{fg,out})$.

The energy balance constraints to be included as equality constraints in the optimization problem given by Equation 4 are as follows:

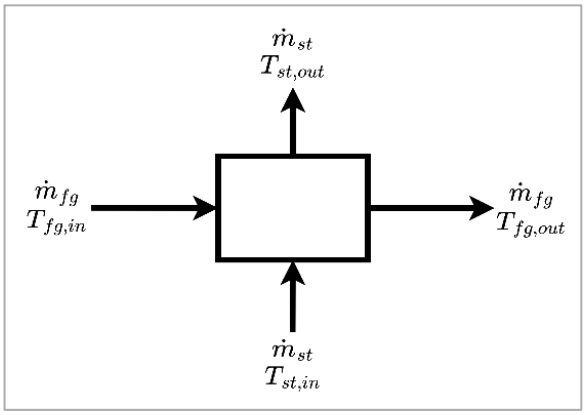

Figure 5. Schematic of lumped superheater system.

$$\dot{m}_{fg}(k)\, C_{p_{fg}}(k)\, \left(T_{fg,in}(k) - T_{fg,out}(k)\right) = \dot{m}_{st}(k)\, \left(h_{st,out}(P_{st}, T_{st,out}, k) - h_{st,in}(P_{st}, T_{st,in}, k)\right) \quad (8)$$

Here, it is assumed that the flue gas is treated as an ideal gas with constant specific heat capacity while specific enthalpies for steam are computed by using the IAPWS R7-97 correlations. Similar to the first case study, model training is done by using the noisy data with constant bias. Plots comparing results from the models built with and without energy constraints are shown in Figure 6.

Figures 6(a-b) show the results from the model built by using the algorithm without energy constraints while Figures 6(c-d) compare the results from the model built by using the algorithm with energy constraints. Figure 6(c-e) show that although the algorithm with energy constraints included leads to biased estimate with respect to the data, it has practically negligible bias with respect to the truth and also satisfies the energy conservation constraints. This is reverse for the model built by using the algorithm without energy constraints.

The key performance measures are shown in Table 2. It is observed that while the RMSE for both output variables have increased when trained using the data with bias for both the models built with or without considering energy constraints, the relative increase is higher for the model built without considering energy constraints. Particularly for the flue gas temperature ($T_{fg,out}$) there is considerable difference in the results obtained by using the algorithm with energy constraints.

Table 2. Performance measures of the models for the adiabatic superheater system.

Mass constrained?	No. of parameters	% RMSE	
		$T_{st,out}$	$T_{fg,out}$
Trained Using True Data			
No	10	1.35	1.98
Yes		0.99	0.23
Trained Using Noisy Data with Bias			
No	8	4.10	4.31
Yes		2.61	0.75

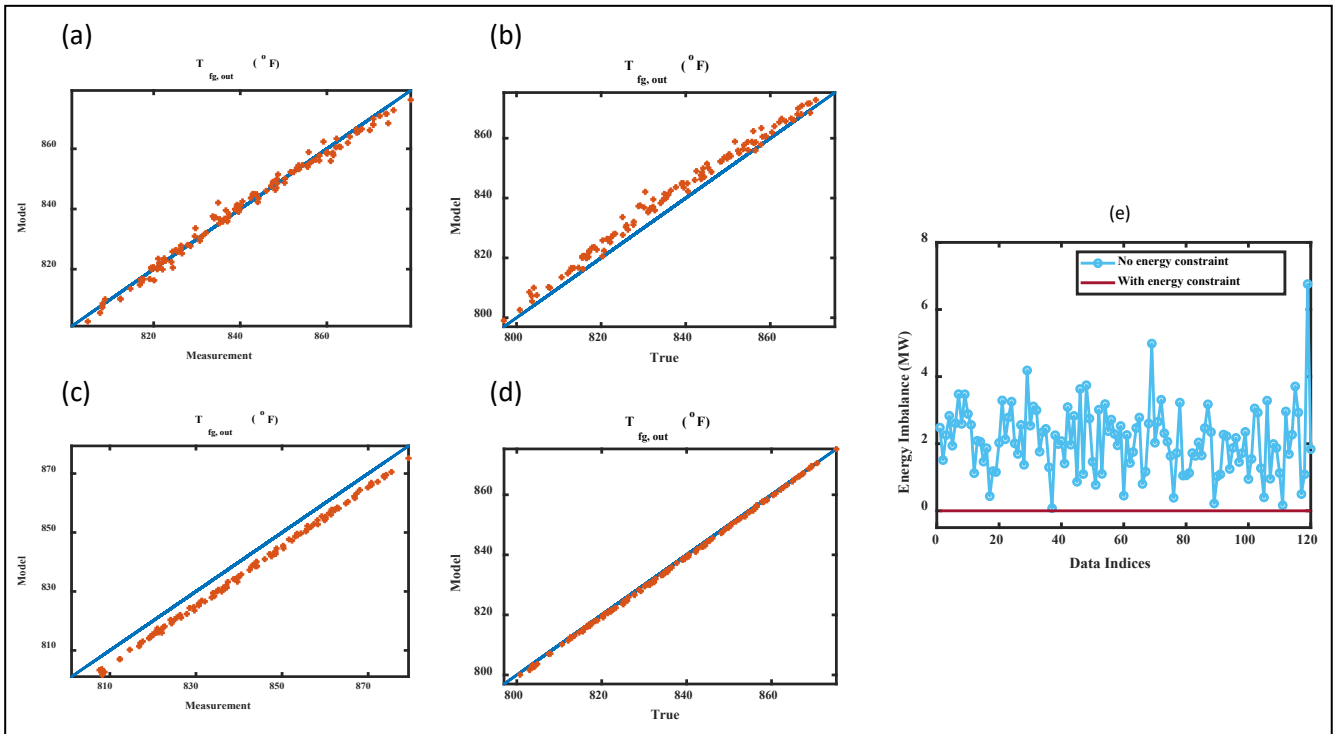

Figure 6. Validation plots for adiabatic superheater model: (a)-(b) BIDSAM, (c)-(d) BIDSAM with energy constraint.

CONCLUSIONS

An algorithm has been developed that guarantees the satisfaction of conservation laws by sparse models obtained by using the enhanced BIDSAM algorithm. The results for the case studies show that the sparse models built by using the algorithm guarantee the satisfaction of mass/energy balance even when trained by using data that are noisy with constant bias with respect to the truth. For the isothermal CSTR case study, %RMSE of the models built with and without mass constraints remains similar when the training is done by using the true data but there was considerable difference in results from the models when training is done by using the noisy data with bias. Similar observations are made for the adiabatic superheater case study as well, there was reduction in the model error with the constraints where energy constraints are satisfied by the enhanced BIDSAM algorithm. One limitation of the proposed algorithm is in the relatively larger training time for the selection of appropriate basis function using the BIDSAM approach. Detailed analysis of this can be found in our previous work [9]. In the future, the algorithm will be extended to application for dynamic model building and for simultaneous conservation of mass and energy while also improving on the computational efficiency for model selection.

ACKNOWLEDGEMENTS

The work was funded by the U.S DOE through the project titled “Boiler Health Monitoring Using a Hybrid First Principles-Artificial Intelligence Model” (Grant # DE-FE0031768). The DOE financial support is gratefully acknowledged by the authors of this paper.

REFERENCES

1. McBride K, Sundmacher K. Overview of Surrogate Modeling in Chemical Process Engineering. *Chemie-Ingenieur-Technik* 91:228–239 (2019).
2. Gonçalves GFN, Assen B, Yuyi L, Lachlan RM, Indranil P, Omar KM. Data-driven surrogate modeling and benchmarking for process equipment. *Data-Centric Eng*. 1, (2020).
3. Panahi M, Khezri V, Yasari E, Skogestad S. Application of surrogate models as an alternative to process simulation for implementation of the self-optimizing control procedure on large-scale process plants-a natural gas-to-liquids (gtl) case study. *Ind. Eng. Chem. Res.* 60:4919–4929 (2021).
4. Brunton SL, Proctor JL, Kutz JN, Bialek W. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U. S. A*. 113:3932–3937 (2016).
5. Cozad A, Sahinidis NV, Miller DC. A combined first-principles and data-driven approach to model building. *Comput. Chem. Eng*. 73:116–127 (2015).
6. Sun W, Braatz, RD. ALVEN : Algebraic learning via elastic net for static and dynamic nonlinear model identification. *Comput. Chem. Eng.* 143:107103 (2020).
7. Fasel U, Kutz JN, Brunton BW, Brunton SL. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 478, (2022).
8. Kaheman K, Kutz JN, Brunton SL. SINDy-PI: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics: SINDy-PI. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 476, (2020).
9. Adeyemo S, Bhattacharyya D. Optimal nonlinear dynamic sparse model selection and Bayesian parameter estimation for nonlinear systems. *Comput. Chem. Eng*. 180: 108502 (2024).
10. Karniadakis GE. Kevrefidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nat. Rev. Phys*. 3:422–440 (2021).
11. Sansana J, Joswaik MN, Castillo I, Wang Z, Rendall R, Chiang LH, Reis MS. Recent trends on hybrid modeling for Industry 4.0. *Comput. Chem. Eng*. 151:107365 (2021).
12. Bradley W, Kim J, Kilwein Z, Blakely L, Eydenberg M, Jalvin J, Laird C, Boukouvala, F. Perspectives on the integration between first-principles and data-driven modeling. *Comput. Chem. Eng*. 166: 107898 (2022).
13. Wächter A, & Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* vol. 106 (2006).

© 2024 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See https://creativecommons.org/licenses/by-sa/4.0/

